Eötvös Loránd University of Sciences
Fakulty of Humanities

THESES OF DOCTORAL DISSERTATION

Szegedi Zoltán

Forensic text linguistics

Doctoral School of Linguistics

Dr. Tolcsvai Nagy Gábor MHAS

Doctoral Program of Hungarian Linguistics
Sociolingvistics Subprogram
Dr. Bodó Csanád PhD.

members of comission:

chairperson:
Dr. Balázs Géza dr. univ., CSc,

official examiners:
Dr. Gallasy Magdolna CSc.
Dr. Laczkó Krisztina PhD.

secretary of comission:
Dr. Sárosi Zsófia PhD.

other members of comission:
Dr. Szili Katalin CSc.
Dr. Zelliger Erzsébet CSc.
Dr. Vörös Ottó CSc.

Supervisor:

Dr. Kiss Jenő, MHAS, professor emeritus

Budapest, 2018

Contents:

1. Introduction, objectives

Forensic linguistics is a young branch of applied linguistics that deals with the investigation of incriminated texts. It primarily examines language usage in texts related to criminal matters but it also has legal issues such as the meaning of the text of the law or the questions of translation and interpretation in the judiciary. Language is also a tool for police and judicial work, and in criminal proceedings it answers questions that can be investigated based on language use and linguistic methods in a particular criminal code. Criminalistic textwriting is characterized by a high level of interdisciplinarity, closely related to criminology and criminal and personality psychology, and within sociolinguistics, psycholinguistics and stylistics within linguistics. Phonetics and phonology are important in speech recognition. While every language user uses the same set of elements and the same rules, all users of the language may assume that they use the language differently than other members of the speaking community. Similar language use situations are similar to the linguistic behavior of the compiler. In his language, many things are not conscious, automated, and despite his will, he also provides information about himself, based on which he can draw conclusions from the author's gender, age, education, main motivation, and personality psychology through his personality traits. With my thesis I wanted to highlight the importance and the timeliness of the subject, the unfavorable changes in the world suggest the appreciation of the area, and in Hungary it has not evolved criminalistical textlinguistics as it is in some other countries. My aim is to show that the methods applied abroad are independent from the language in which the text to be tested is written, so that they can be used at home.

2. Materials, hypotheses, methods, goals

The first statement – that every language user has his own unique code – is the starting point and main hypothesis of my dissertation. Ferenc Nagy's categories can be expanded and more about the writer than gender, age, occupation and culture. The signs of Tuhn's communication style and Correll's core motivation are found in most texts. There are signs in the texts that make the writers distinct from each other, and those that make it possible to find that texts come from the same author. Such signs are language habits, errors, text phenotype. Despite the fact that the texts are very similar in nature because of their subject matter, function, and material used, there are no two identical, unique features.

My method, as in the case of criminalist text-language studies, is customary to examine the material at all levels of the language: the phonemes, the morphs, the lexemas, the syntax and the text. Of course, there are not always any usable signs at all levels. The content or difference between the texts and the text phenotype helps to determine whether texts come from the same or different drafters.

I have based the general characteristics of the anonymous and pseudonym letters on 40 texts.

The short, instant messages come from communication with 8 different recipients, their length ranging from 9 to 15 sentences, the number of messages being surveyed is around 500.

The other test material came from computer correspondence close to chattering. The recipients and the drafters were also different people, but during the scan, I was able to contact the letters to some senders who sent messages to various recipients of various nicknames. The number of very different leaflets far exceeded 500.

3. Structure of the dissertation

In Chapter 1 of my dissertation, I defined the field of research in criminal linguistics, I placed it within linguistics, and I named the linguistics and other sciences related to it.

In Chapter 2, I present the hypotheses and methods summarized above.

Chapter 3 deals with the history of criminal linguistics, with particular reference to publications in Hungary, from the beginning to the present.

In Chapter 4, I presented the most sociolinguistic components that make up the basis of criminalist text language studies. Social groups and socio-cultural factors influence the language's use of the individual.

Due to linguistic diversity, language users follow different standards. The norms can be recognized in the text, and the conclusion can be drawn to the formulator.

The common language is the language variant that is related to all versions of the language, and deviations from norms in criminalist textual language generally refer to deviations from the common standard.

Dialect phenomena appear rarely in written texts, so the planner's geographic location is usually cumbersome.

The group is characterized by a kind of internal language usage, a specific vocabulary, one of which is to increase the efficiency of intra-group communication and to strengthen group identity. This makes it easier to group your compiler into a group. One of the procedures of

criminalist texting is, in fact, nothing more than the classification of a writer in as many groups.

Individual creativity creates idiolytics: every language user uses the same language slightly differently. This distinguishes the person from other members of the group. This makes it possible to do criminalical textlinguistical investigations.

Slang is no longer just about the language of the youth, but it helps to estimate the age of the writer and highlights the relationship with the language of drafting and the topic of the text, so it is an important sign in criminalist text language studies.

In the thieves slang, we encounter the elements of slang and the language of the prisons, when the language profiles one of the texts of the examined text were created, the lack of these expressions indicated the role of the drafter.

The language of women and men differs slightly from one another. Although the differences between women's and men's social roles have resulted in a decrease in language differences, the majority of authors who speak of the differences between the two genders language use as facts.

There are significant differences between verbal and written texts. The changes that have taken place over the last decades have resulted in certain written texts showing the characteristics of verbal texts, and the limit of verbalism and literacy is obscured. Typically, the material of criminological text-language investigations are written texts, among which chapters and e-mail are also found more and more often. Digilect can help identify the author. Selecting a communication channel already reveals something about the sender itself, and the channel also determines certain features of the text.

The most varied area of the language is the vocabulary. There have been also significant changes in this area in a short time. The changes in today's Hungarian language for the texts of criminal linguistics are particularly important because they can be of information important for the estimation of the author's age.

Chapter 5 of my thesis deals with style. Styling is one of the most important parts of the criminal investigation. Style selection between different options, at different levels of a language. The multitude and combination of linguistic elements puts the composer in a constant choice, who chooses to communicate in a manner appropriate to the speech situation. You can also rely on patterns, these types of text help you to achieve your goals. Style unfolds in language usage, perceives others and adds extra meaning to the text. However, the above choice between elements and options is not always conscious. Certain decisions of the

language user are automated. They are regular and can not be manipulated so they are useful in criminalist text language studies.

In different speech situations, we follow patterns and traditions. Non-language factors, language variants define text types that have a distinctive style. This is the type of style. However, this does not mean that the user does not have room for maneuver. You can even cross the boundaries of tradition. Certain features, however, do not come from individual preferences but are inherent in text types and style types. The latter two facts are extremely important in criminalist text language studies.

Style is the result of deliberate and unconscious decisions of the language user, and an important means of language self-representation. Idiotism is a multitude of features that appear in linguistic behavior.

Two types of style examination are qualitative and quantitative. The quantitative analysis is based on a statistical process, while the qualitative style analysis is not the frequency, but the notion of style notes.

Chapter 6 of the dissertation deals with the errors. Analysis of errors is another important tool in the forensic linguistic study, where valuable information can be obtained. A distinction must be made between the competence and the performance error. From the point of view of the tests, the former are important because they result from the language user being unaware of the rule or misapplying it. The reason for the language mistake may be the transfer when the mother language affects the foreign language. The frequency of spelling mistakes in the text can be deduced from the formatting of the writer.

In the case of incriminated texts, the draftsman intentionally creates an error in the text. This is done with the intention of disguising himself or leaving false traces behind him. However, this is imperfectly accomplished.

In Chapter 7, I listed the striking marks that may be of relevance to the test, but are not mistakes. Among other things, the external form of the text and the text structure.

In Chapter 8, I deal with the content character of the incriminated texts. Determining the truth of the text is just a matter of other sciences, but we also find linguistic signs that point to what really happened or fictitious.

Signs of experience are linguistic signs, such as the detail richness of the text. The list of credibility criteria for the text can make a good help of judging whether there is a true experience in the background of a text. It is also important to have a good form of text. Lying is causing psychic pressure, resulting in unambiguous logic, inconsistency, and contradictions in the text.

In Chapter 9, I was dealing with language and use of language. The basic unit of communication is the text. Different texts from the same author can be very different, while texts from different authors may be similar. This is a role for text types.

The reason for the use of texttypes for communication is the pursuit of efficiency and economy, and it also means constraints.

The blackmail letter is, according to some people, a type of text, others do not think this, but there are of course characteristic content- and form elements. They are the way of claim, threat, and claim fulfillment. Based on its function and thematic pattern, it is distinct from other texts. The writer sometimes uses known types of texts, leaf patterns, in which case the blackmail letter shows the characteristics of the business or official letter. The use of language elements and rules also provides a point of reference as to how well the writer has the ability to create written texts and the degree of his or her level of education.

Sometimes we may encounter signs of language courtesy in blackmail letters. This is an explicit rational behavior that increases the chances of success.

Each function is assigned a function. The basic function of the text is the author's intention. The basic function of the blackmail letter is derived from a speech, blackmail. This also includes information and a contact function. In addition to the mandatory elements, there are some parts that do not necessarily appear in the blackmail letter. There is a transition between extortion and threat.

Chapter 10 shows the phenomenon of blackmail. Often, this is the first offense committed by a person because it is easy to commit. It has three sections contact, negotiation and handover.

In chapter 11, I was dealing with the author, pseudonym and anonymity. The author is the person who creates the text. The author of the article is not the same as the one who writes the text in writing, and there is an example of collective authorship. The alias sometimes carries information about its wearer. Since it is a chosen name, it is no coincidence of who the alias is chosen.

Chapter 12 deals with science, criminology, psychology and computer linguistics related to criminal linguistics, and the cooperation of the three sciences. I have presented the working groups and methods abroad.

Computer linguistics in the future can do a lot for criminological texts, but in the present it is being heavily supported by various programs. One of these is the Ellegård test, the essence of which is to characterize the individual characteristics of the original text in question with the characteristics of the written texts of the suspect. Distinctive features should be systematic and statistically independent from each other. Such occurrences reduce the number of

suspects by statistical and combination rules. There are at least two distinct features of different language levels which, if no match is found in the test, are likely to come from different authors.

KISTE is a special corpus made up of crimson texts which, besides research, serve as a comparison model in the context of authoring and linguistic profiling. Similar would be needed in Hungary.

Concordance analysis is a computer-aided review procedure. The program lists all the words of the texts to be compared in alphabetical order, thus determining the frequency with which the texts occur.

Koppel succeeded in automated profiling. Its method is based on a statistical process, and the computer-generated profile includes age, gender, and whether the speaker is English-native.

Title of Chapter 13 Authentication. Authentication involves two things: it means finding out who has written the text in question and answering the question of whether a particular text may originate from a specific person.

In this chapter, I wrote about collective authorship and summarized the axioms of the linguistic examination of written texts from an unknown author.

Language profiling is nothing other than categorizing the author. Regarding the gender of the author, the majority of women are still treated as feminine bribes. There is a changing picture about obscene and that women are less obscene. The type of language offense also refers to the offender's gender.

The communication channel chosen tells you something about the author's age. During language profiling, it is possible to estimate age roughly. The processes in the language give this point.

From the quantity and quality of the errors, the size of the vocabulary can be deduced from the educator's educational level and the literacy.

There are seldom terms that refer to the author's work or specific knowledge.

The underlying motive has five categories that can be inferred from language use. Texts are linguistic products that can be inferred from the group-specific attributes of the formulator, in addition to those listed above, for their motivation.

The Language and Behaviour Profiles (LAB) is based on the unconsciousness of personality structures in language use. In addition to motivation, the pattern of reaction is also important, and the author's future behavior can be predicted.

The author's style of communication is recognizable in the text. Based on the communication style, we can deduce some of the other attributes of the draftsman, as well as the basic feature of his personality.

In this chapter I presented the language profiling with two anonymous letters, including Ferenc Nagy categories, as listed here.

I also explained the findings that I have collected during the examination of anonymous letters.

Sequencing is performed when the investigation is fired for some reason. In this case, the short details and clauses of the text are examined one by one, always emphasizing the expected continuation and the reason why the text does not continue as expected.

In this chapter, I also present the tests whose materials were provided by correspondence on a web page. Firstly, I found that the letters originated from the same author, and based on thematic, language-use and text phenotype, I could link other letters to the author even though their sender used another nickname. The same and similar sentences, structures, mistakes made this all possible. The process repeated several times, behind the five nicknames the same person was. From the focus of his letters he could also deduce what motivates his goals. In one of the letter cycles, the sender tried to release a teenage boy, but some of his expressions did not correspond to the role, and other language-specific features were assigned to the other four letter cycles. Our language behavior was automated to such an extent that we did not intentionally change it altogether.

Two other nicknames also showed similarities in content and language usage. The investigation revealed that a person uses the two nicknames, but this person is not the same as the one of the five leaflet formulas described above. There was another motivator of the draftsman, others were the language features and mistakes. Motivation can sometimes be deduced from explicit data, sometimes focus, and can be used to identify or differentiate people.

Title of Chapter 14 of the dissertation is Examination. In this chapter, I wrote about the race of criminals and law enforcement agencies. Technical changes favored the spread of certain types of crime. As we have seen, it is easy to get into the confidence of young people during chat, and the means of communication are also exploited by terrorist organizations. In response to linguistic profiling, an antidote is launched, a program that pays attention to individual language features and offers the option of masking them by using other terms.

Network research is not only interesting in the field of sociolinguistics. It can also be valuable in investigating terrorist organizations and organized crime groups.

Criminalistical textlingusistics is unimaginable in the future without computer aid. It would be necessary to create a criminal intelligence database and to expand the LAB categories through personality psychology and communication psychology.
In Chapter 15 I summarized the above.

4. Results, Theses, conclusions

The first part of the dissertation contains the methods and methods developed in the last decades. I present the domestic history of the field of science, the most important works and the foreign results.
When examining anonymous mails, I came to the conclusion that dialect expressions are hardly present in texts, so it is not possible to use the definition of the author's geographical location. Indicators of age appear in larger numbers in the letters, but generally speaking of the draftsman is that it is young, middle-aged or old. Traditional gender roles changed, but differences in language use were also declining, but most of the messages contained some sign of gender. These signs are not always found at the level of lexems. As vocabulary is the most sensitive to social change, the omission of the boundaries of social roles has a significant impact on word-use. Focus is given to target or problem orientation. Signs of education and education can be recognized in the texts with greater certainty, as is often the case with Tuhn's communication style and Correll's motivation. The emotional state of the writers can be inferred in half of the cases, even if explicitly expressed by a few.

full corpus: indicators	**40 anonymous letters appearance number**
mother language:	40 hungarian
a sign of distortion of language usage:	up to 4, at the spelling level
linguistic socialization or lokation:	in 4 cases dialect phenomenon
Age determination:	26 cases
old style:	3
slang:	20–22
determination of sex:	35 cases
male writer:	17

female wirter:	18
determination of education:	35 cases
emotional state:	19 cases (male writer 7, female writer 12)
Correll's motivation, Tuhn's communication style	33

I have presented the authorship test with messages from a rapporteur with eight different people. In addition to content identity, I found recurrent spelling mistakes, recurring expressions, and features of the phenotype of sentences that made it very clear that the writer was always the same person. Content Identity in this case meant that in the conversations with different participants, the writer named the same garments and always ordered the same amounts for the photos in the garments. Some misplaced words repeated in a message to several recipients, as there were several examples of leaving the last letter of the sticker.

During linguistic profiling, I was able to expand my categories, besides the categories of gender, age, education and occupation, the number of authors, the identity of the author and the recording person, the language distortion, the mother tongue and the language of socialization, the age, education, language level, occupation, textual practice, Correll's motivation and Tuhn's communication style, the author's emotional state, motivation, and dangers were included in the table of my categories. This is a significant expansion compared to the categories of Nagy, even if it is not always possible to enter something in each column of the table. More information is an advantage in both the reconnaissance and the investigation and the trial phase.

The grouping of large numbers of letters and authors is a serious result, as is the separation of texts of similar topic, based on language use and motivation. It has made hard for the JT code sender to send a letter to 37 people, but the identity of the author is not a reliable starting point, since the users frequently change their nickname on the website, which has different causes and which I mentioned in my dissertation. Finally, however, it was verified that these letters were probably written by the same person. In another letter cycle, the nickname was used by the draftsman, his messages, 23 letters written to a user. During of the investigation, it was possible to ascertain that the texts are likely to be the same as the original starting sentences, repeated expressions, frequent words, constant typing of spelling, references to extraterrestrial reality, the same type of emoticons, the person who wrote the letters

mentioned above. When the correspondence ran to the reef, the compiler started to correspond with the recipient of the nickname, PG, to himself as a teenage boy, trying to reach the trust of the teenager. Thus, a new correspondence cycle started, which initially showed deviations in the punctuation compared to the previous speaker's language. Soon there was a sentence that did not feature at all in the language of the age group. In the second half of the letter cycle, they appear in the PG's letters of language used by the former draftsperson. The number of punctuation decreases, the thematic similarity is displayed, and the marking error of previously repeated words and the duration of the vowel is displayed. The composer played a role here, but he was not able to play it credibly and consistently. Later, two other nicknames were sent to the letter cycle, with the same language usage features being regularly found in the aforementioned writings. The consignments sent by a formulator (M) from five usernames are summarized below:

	JT	JT2	PG	SZM	SZM2
subject-matter correspondence	+	+	+	+	+
lexical correspondence	+	+	+	+	+
emoticon	+	+	+	+	+
a punctuation other than usual	+	+	+	+	+
incorrect vowel time	*segit, szolit, eltüntek, irtad, tüntél, állit, tágitás*	*könnyü, szinű, kisérő, ird, tágitás*	*suly, viz, megirta, elötte papir, mult, amiota, huzzam,*	*irta, segiteni, ugy, gunyolodnak, leirom, fiu, irták,*	*irj, vizben, fiu, okosit, segit, igy, tul, megtanitalak*
incorrect spelling of preverb and verb	*eltudjátok mondani*	*megszoktam kérdezni*	*megakartam mondani*	*lemerem irni*	
greeting	*szia* (rarely)	*szia* (rarely)	*szia* (rarely)	*szevasz, szeva, csá, szia*	*szevasz, szeva, csá, szia*
signature, reference to himself	no sign or direkt reference *M* (1. forename)	no sign, no reference in text	no sign, no reference in text	*M* (forename)	*M* (forename)

The identity of LZ and SF nicknames is characterized by thematic similarity, signs of language courtesy, long vowels in short letters, errors in spelling use, and abbreviations. However, this person had a different motivation, and other differences also indicated that he was not the same person (M) whose correspondents were discussed above.

Signs of identity in texts from SF and LZ

	SF	LZ
question formulation	*mennyi idős vagy? hogy nézel ki? haj szem suly magasság*	*mennyi idős vagy? hogy nézel ki? haj szem suly magasság*
language courtesy	*esetleg meg tudnád mutatni, tudnánk találkozni, fel tudnánk dobni*	*tudnánk találkozni, tudnál küldeni*
incorrect vowel time	+	+
use punctuation	does not capitalize, the commas are inconsistent, instead of using an emoticon, emoticon, question mark cumulation	does not capitalize, the commas are inconsistent, instead of using an emoticon, emoticon, question mark cumulation
motivation wording of the intent of meeting	showing genital organs in camera, meeting, sexual relationship „*neked lenne kedved 1x találkozni?”*	taking a picture of the genitals, meeting, examination” of genitals „*titokban akkor tudnánk találkozni”*
intent of changeing the communication channel	+	+
abridgment	*lax, szal* (2), *mizujs* (2) *am* (8)	*lax* (2), *szal* (2) *am* (6)
preverb usus	*feltudnám dobni, megutdod mutatni stb.*	*megtudjam vizsgálni*
expressing consent	*oksa, oksi, oké*	*oksa, oké*

specifying of penis	*fütyi*	*fütyi*

Similarities and differences between the textes from LZ (=SF) and M

	similarity and consistency	differences
topic	health advice	
motivation		M: correspondence, sometimes request of a photo LZ: request of a photo, meeting
words	*huzni, vizbe, segitett, emailcim, mig*	M: *pöcs, Hány éves vagy?* LZ: *fütyi, kuki, Mennyi idös vagy?*
spelling	good spelling	
emoticon		M: XD LZ: :) :))))
punctuation		M: word, space, comma, word / word, space, point/?, word; does not cramp punctuation LZ: word, comma, space, word / word, space emoticon, question mark cumulation, does not use a point
abridgment, reduction of the number of		M: reduction of the number of characters,

characters		abridgment not typical, uses capital letters LZ: reduction of the number of characters, abridgment is typical ; doesn't use capital letters
spread of messenges		M: short and long texts too LZ: just short texts,

My research also had an area that aimed at deliberate distortion of language use. In my experiment, I observed the same tendencies as Dern in his previous experiment. Women use a little more words than men, although this difference was not significant in the first half of the experiment, and the number of words used in both sexes decreased significantly when distorting language use. It was also noticed that the participants in the experiment did not typically think about changing their language usage. When instructed to do so, the distortion was poorly implemented. The manipulation mostly affected spelling and vocabulary, reduced style, more slang, but camouflage was inconsistent.

The number of used words in the letters

	average number of words	women	men
1. task (spontaneous)	72	74	70
2. task (distorted)	44	39	49

Based on the above results, the following theses can be formulated:

Thesis 1: The categories of linguistic profiling can be significantly expanded, in addition to gender, age, occupation, and literacy, there are usually signs of the author's number, language

distortion, language usage, text creation, motivation and communication style and emotional status too.

Thesis 2: Intentional distortion of language use can not be performed at a level that would not be traced during the investigation. Changing the use of languages that are automated for everyone does not even think of people with a higher literacy than the average, and performing the task is a great intellectual effort, which they can only partially perform, their attention does not extend to all levels of language, even though they are built on each other.

Thesis 3: Writers' writings point to multiple levels of agreement on the basis of which the writer can be bound to the same writer even if he sends messages to different names for deception or attempts to believe that he is someone else as the one he has previously communicated to the recipient.

Thesis 4: There are also signs that, despite the similarity of the texts by theme or function, texts from different authors can be distinguished from each other.

Thesis 5: Criminalistical text linguistics not only can be used to examine texts in the classical sense, such as letters, to produce results in short messages and chatting if they can be examined in relatively large quantities.

Criminalistical text linguistics can do a lot for public security. Different texts, even texts without greater care, than the sentiments sent during the chat, are capable of getting to know some of the personality traits of the draftsman. In this case, it may even be a good thing for the draftsman to "ignore", that is to say, the natural, the characteristic, the informal use of the language. So you betray yourself more than you would think. The expressions, even if their scope is short, since they are created in large numbers by chatting, are capable of drawing on information about the social characteristics, personality, categorization of the formator, and thus narrowing the scope of possible drafters, or even future their predicaments, and the utterances - even if they were transmitted from different nicknames or even from different forums - to bind to a particular draftsman.